Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome

Omasits, Ulrich ; Quebatte, Maxime ; Stekhoven, Daniel J ; Fortes, Claudia ; Roschitzki, Bernd ; Robinson, Mark D ; Dehio, Christoph ; Ahrens, Christian H

Abstract: Prokaryotes are, due to their moderate complexity, particularly amenable to the comprehensive identification of the protein repertoire expressed under different conditions. We applied a generic strategy to identify a complete expressed prokaryotic proteome, which is based on the analysis of RNA and proteins extracted from matched samples. Saturated transcriptome profiling by RNA-seq provided an endpoint estimate of the protein-coding genes expressed under two conditions which mimic the interaction of Bartonella henselae with its mammalian host. Directed shotgun proteomics experiments were carried out on four subcellular fractions. By specifically targeting proteins which are short, basic, low abundant and membrane localized, we could eliminate their initial under-representation compared to the estimated endpoint. A total of 1,250 proteins were identified with an estimated false discovery rate below 1%. This represents 85% of all distinct annotated proteins and about 90% of the expressed protein-coding genes. Genes that were detected at the transcript but not protein level, were found to be highly enriched in several genomic islands. Furthermore, genes that lacked an ortholog and a functional annotation were not detected at the protein level; these may represent examples of over-prediction in genome annotations. A dramatic membrane proteome re-organization was observed including differential regulation of autotransporters, adhesins and hemin binding proteins. Particularly noteworthy was the complete membrane proteome coverage, which included expression of all members of the VirB/D4 type IV secretion system, a key virulence factor.

DOI: https://doi.org/10.1101/gr.151035.112

Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome

Ulrich Omasits, Maxime Quebatte, Daniel J. Stekhoven, et al.

Genome Res. published online July 22, 2013
Access the most recent version at doi:10.1101/gr.151035.112

Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome

Ulrich Omasits[1,2], Maxime Quebatte[3], Daniel J. Stekhoven[1], Claudia Fortes[4], Bernd Roschitzki[4], Mark D. Robinson[1,5], Christoph Dehio[3], Christian H. Ahrens[1]

[1]*Quantitative Model Organism Proteomics, Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

[2]*Zurich Life Sciences Graduate School Program in Systems Biology, Zurich, Switzerland*

[3]*Biozentrum Basel, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland*

[4]*Functional Genomics Center Zurich, ETH & University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

[5]*SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland*

Corresponding author:
Christian H. Ahrens
Quantitative Model Organism Proteomics, Institute of Molecular Life Sciences
University of Zurich
Winterthurerstr. 190
CH-8057 Zurich, Switzerland
Phone: +41 44 635 3960 E-Mail: christian.ahrens@imls.uzh.ch

ABSTRACT

Prokaryotes are, due to their moderate complexity, particularly amenable to the comprehensive identification of the protein repertoire expressed under different conditions. We applied a generic strategy to identify a complete expressed prokaryotic proteome, which is based on the analysis of RNA and proteins extracted from matched samples. Saturated transcriptome profiling by RNA-seq provided an endpoint estimate of the protein-coding genes expressed under two conditions which mimic the interaction of *Bartonella henselae* with its mammalian host. Directed shotgun proteomics experiments were carried out on four subcellular fractions. By specifically targeting proteins which are short, basic, low abundant and membrane localized, we could eliminate their initial under-representation compared to the estimated endpoint. A total of 1,250 proteins were identified with an estimated false discovery rate below 1%. This represents 85% of all distinct annotated proteins and about 90% of the expressed protein-coding genes. Genes that were detected at the transcript but not protein level, were found to be highly enriched in several genomic islands. Furthermore, genes that lacked an ortholog and a functional annotation were not detected at the protein level; these may represent examples of over-prediction in genome annotations. A dramatic membrane proteome re-organization was observed including differential regulation of autotransporters, adhesins and hemin binding proteins. Particularly noteworthy was the complete membrane proteome coverage, which included expression of all members of the VirB/D4 type IV secretion system, a key virulence factor.

INTRODUCTION

A major goal of the post-genome era is to understand how expression of the functional elements encoded by a genome is orchestrated to allow an organism to develop and adapt to life under varying conditions. Transcriptomics and proteomics technologies both provide important and complementary insights: the former allow researchers to generate global quantitative gene expression profiles and to study gene regulatory aspects like the impact of short RNAs. However, due to the varying correlation of transcriptomics and proteomics data reported in the literature (de Godoy et al. 2008; de Sousa Abreu et al. 2009; Maier et al. 2011; Marguerat et al. 2012), the direct measurement of protein expression levels is often desirable. For certain aspects, proteomics data can provide more informative and accurate data, as it reflects the effects of other important regulatory processes like protein translation rates and protein stability (Schwanhausser et al. 2011). Furthermore, proteomics provides unique functional insights including post-translational modifications, subcellular localization information and identification of interaction partners of proteins.

Due to enormous advances in mass spectrometry instrumentation, biochemical fractionation methods and computational approaches, proteomics has matured into a state where the description of complete proteomes expressed in a specific condition is within reach. So far, only one study has claimed the identification of a complete proteome expressed in haploid and diploid baker's yeast (de Godoy et al. 2008), while extensive proteome coverage has been reported for several prokaryotes (Becher et al. 2009; Jaffe et al. 2004; Malmstrom et al. 2009) and archaea (Giannone et al. 2011). Describing extensive proteome maps under different conditions with a discovery proteomics approach is an important first step in defining the protein expression landscape for an organism and facilitates a subsequent shift away from the discovery mode to a re-measurement or scoring mode (Ahrens et al. 2010; Kuster et al. 2005).

Due to the lower transcriptome and proteome complexity compared to eukaryotes, an exhaustive discovery proteomics approach is particularly amenable for prokaryotes. We describe here a generic strategy to achieve an essentially complete coverage of a prokaryotic proteome expressed under specific conditions. Key elements of the strategy are the parallel extraction of RNA and protein from matched samples, and a saturated transcriptome analysis by RNA-seq (Wang et al. 2009). This in turn allows to generate a condition-specific endpoint estimate of the number of actively transcribed protein-coding genes, which is a more appropriate estimate than considering all annotated protein-coding genes. A combination of experimental and computational strategies is then used to dig very deep into the proteome.

We apply the strategy to two conditions that mimic the changing environment encountered by *Bartonella henselae* upon transfer by its arthropod vector into its mammalian host. The Gram-negative α-proteobacterium *B. henselae* is a hemotropic, zoonotic pathogen that frequently causes cat scratch disease in immuno-competent humans, as well as bacteraemia, endocarditis, and vasoproliferative lesions in immuno-compromised patients. Members of the genus *Bartonella* are considered re-emerging pathogens and are primarily being studied as models for host-pathogen interaction (Harms and Dehio 2012). A particular emphasis was put on achieving an extensive coverage of the important membrane proteome (Savas et al. 2011). Membrane proteins carry out essential functions as transporters, enzymes, receptors to sense and transmit signals, and adhesion molecules. In light of the resurgence of infectious diseases, membrane proteins are furthermore prime candidates for the development of urgently needed novel anti-infectives (Norrby et al. 2005).

Relying on a very stringent false discovery rate (FDR) cutoff, we were able to identify 1,250 of the 1,467 annotated distinct *B. henselae* proteins, i.e. a proteome coverage of 85%.

Several lines of evidence indicated that we have exhaustively measured the expressed proteome and can claim to have identified a complete membrane proteome. This included expression evidence – to our knowledge for the first time – for all protein components of a bacterial type IV secretion system (T4SS) which spans the inner and outer bacterial cell membranes.

RESULTS AND DISCUSSION

Model system to explore complete proteome coverage

We chose *B. henselae* as a model system due to several reasons: i) its relatively small genome (1.93 Mbp) comprises 1,488 predicted protein-coding genes (Alsmark et al. 2004); ii) it is a facultative intracellular pathogen that can be grown in pure culture; iii) protocols for subcellular fractionation have been described (Rhomberg et al. 2004); and iv) *in vitro* conditions that mimic the pH dependent induction of virulence genes required for the successful interaction with host endothelial cells, the likely primary niche for *B. henselae* (Harms and Dehio 2012), have been established (Quebatte et al. 2010). The availability of a model system that eliminates the need for co-culture with human endothelial cells is critical to achieve complete coverage of an expressed proteome.

Our *in vitro* model system relies on the induction of the transcription factor BatR (BH00620) that is essential for the pathogenicity of *B. henselae* (Quebatte et al. 2010) (for details, see Supplementary Methods, Tables S1, S2). In the absence of IPTG (uninduced condition), the BatR regulon is not induced, resembling the situation encountered in the arthropod midgut. In contrast, BatR expression is up-regulated in the induced condition, resulting in a marked induction of the BatR regulon, including the VirB/D4 type IV secretion system (T4SS), which is required for infection of endothelial cells (Schulein and Dehio 2002). This state mimics the environment encountered by bacteria in the mammalian host.

A generic strategy for complete proteome coverage by discovery proteomics

We rely on our previous definition of complete proteome coverage, i.e. having identified protein expression evidence for the annotated protein-coding genes actively transcribed in a given state (Ahrens et al. 2010). A recent proteogenomics study of 46 prokaryotes indicated that on average only 0.4% protein-coding genes were missed in the original genome

annotations (Venter et al. 2011), justifying our focus on the reference genome. Our strategy to achieve an as complete as possible coverage of the expressed proteome of a prokaryote consists of three stages.

In a first stage, RNA and proteins are extracted from identical samples, and whole transcriptome libraries are sequenced to saturation by RNA-seq (Figure 1A). Thereby, the number of protein-coding genes actively transcribed in a given state can be estimated, shown here for the sum of protein-coding genes expressed in the uninduced and induced condition (orange dashed line, Figure 1B). Based on such an optimal endpoint estimate, in a second stage several pilot experiments are performed on cytoplasmic and total membrane fractions of the respective conditions. Following a statistical comparison of the pilot phase proteome (green line, Figure 1B) to the predicted endpoint, areas of under-representation can be targeted by the analysis-driven experimentation (ADE) feedback-loop strategy (Brunner et al. 2007), which can help to overcome the premature saturation of distinct protein identifications and sequence deeper into the expressed proteome (blue lines, Figure 1B). In a third stage, evidence is presented that virtually no biases remain when comparing protein parameters of all identified proteins to those called actively expressed, justifying the claim to have identified a complete proteome expressed in a specific condition. Analysis of such a dataset is expected to provide novel insights regarding the achievable membrane proteome coverage, differential protein expression, and evolutionary conservation and genome structure (Figure 1C).

Transcriptome exploration by RNA-seq

We relied on RNA-seq (Wang et al. 2009) primarily to generate an endpoint estimate for the number of expressed protein-coding genes. Whole transcriptome libraries of two biological replicates per condition were generated using a protocol that enriches for mRNA transcripts (see Methods). We sequenced very deep into the transcriptome and obtained 55-87 million

single end 50-mer reads per sample. Of these, 10.7-26.7 million reads mapped unambiguously, while the vast majority of remaining reads originated from multiple-copy rRNA genes (see Methods, Table S3). RPKM values (Mortazavi et al. 2008) showed very high concordance of the biological replicates ($r > 0.97$, Figure S1).

To estimate how many protein-coding genes are actively expressed in the two conditions, we plotted the number of distinct expressed protein-coding ORFs as a function of the sum of uniquely mapping reads. We required at least 5 distinct reads within a 50 nt window of the 5' end to deem a protein-coding gene actively expressed (Figure S2), a cutoff similar to that used by (Wang et al. 2009). Saturation is characterized graphically through flattening of the curves as the number of reads increases. Due to the asymptotic nature of saturation curves, reaching complete coverage is theoretically only possible with infinite effort. Therefore, we define saturation as the number of discoveries from where, based on nonlinear modeling and extrapolation, a doubling of effort is expected to increase the number of discoveries only marginally. Figure 2A indicates that doubling the number of reads would increase the number of detected protein-coding genes by less than 3.5% for sample uninduced2 and by roughly 1% for induced2. Therefore, our analysis indicated that the transcriptome was sequenced to saturation (Figure 2A). We acknowledge that different library preparations might potentially identify additional genes, and that very low abundance transcripts (and proteins) expressed in only few cells of the population may not be identified with this approach.

We also plotted the density of the RPKM values in order to assess the distribution of transcription levels for all annotated protein-coding genes: the resulting bimodal graph suggested that under the conditions studied not all protein-coding genes are actively expressed; RPKM=10 might be considered a conservative lower cutoff (Figure 2B). The average RPKM values for members of the *virB*/*D4* operon in condition uninduced2 (30),

where the operon is expected to be expressed at low levels, versus induced2 (160) support this observation.

Based on the combined thresholds 1,353 protein-coding genes were expressed in the two conditions (uninduced 1,254 and induced 1,349). An inter-replicate analysis revealed more than 95% overlap of the expressed protein-coding genes (Table S4). We include an error envelope of ±2.5% to account for uncertainty in the thresholds (Figure 3A).

Extended proteome coverage strategy: experimental and computational approaches

Our experimental strategy to reach very deep into the proteome relied on four elements: First, we used a combination of sub-cellular fractionation and additional biochemical fractionation regimens to reduce the overall sample complexity, a measure that had been key to describe the complete expressed proteome of baker's yeast (de Godoy et al. 2008). Second, an exclusion list approach (Kristensen et al. 2004) was applied, which helped to identify a significant amount of low abundant proteins (Figure S5). Third, we relied on the analysis-driven experimentation (ADE) feedback-loop strategy (Brunner et al. 2007) (Figure 1B) to target under-represented areas of the proteome and overcome premature saturation. Finally, for all membrane-derived fractions we used chymotrypsin in addition to trypsin, thereby maximizing the per-protein sequence coverage and the overall membrane proteome coverage (Fischer et al. 2006).

In terms of computational approaches, we combined results from two database search engines, Mascot-Percolator (Brosch et al. 2009) and MS-GF+, an updated version of MS-GFDB (Kim et al. 2010) (see Methods), which employs the generating function approach (Kim et al. 2008) to compute statistical significance of peptide identifications (spectral probabilities). Based on these spectral probabilities or the target-decoy option, one can estimate and stringently control the FDR rate, a critical step for a complete proteome discovery project. Otherwise, lower quality PSMs will start to accumulate false-positive

peptide evidence for proteins in a random fashion (Reiter et al. 2009). In addition, the error propagates and increases from spectra to peptides and proteins (Nesvizhskii 2010); a PSM level FDR of 1% can correspond to a protein level FDR of 8-11% (Balgley et al. 2007). We therefore chose a very stringent PSM FDR cutoff of 0.01%, allowing us to report protein identifications with an FDR below 1% (see below).

Identification of the complete expressed *B. henselae* proteome

The induction of *batR* and *virB/D4* T4SS expression was more pronounced for the sample pair uninduced2/induced2 than for its biological replicate based on the RNA-seq data. Subcellular fractions from this sample pair (i.e. cytoplasmic (Cyt), total membrane (TM), inner (IM) and outer membrane (OM) fractions) were thus analyzed in detail using different biochemical fractionations (see Methods, Figure 1A).

We first measured the Cyt and TM fractions of both conditions using OFFGEL electrophoresis at the protein level (OGEprot). When requiring at least two independent PSMs to identify a protein, 924 distinct proteins were identified in four experiments, i.e. 63% of all 1,467 distinct annotated proteins or 68%±2% compared to the RNA-seq endpoint estimate of 1,353±34 expressed proteins (Figure 3A). Analysis of the IM fractions from uninduced and induced condition ($IM_{u/i}$) and the $OM_{u/i}$ fractions contributed 130,000 additional PSMs (72% more PSMs), but only added 22 previously not identified proteins (Figure 3A), indicating that we were already in the saturation phase. We fitted a saturation curve to the eight OGEprot experiments, which shows the anticipated trend of further protein identifications assuming no change in the experimental approach, and also calculated confidence intervals (see Methods, Figure 3A). Carrying out further OGEprot experiments is predicted to lead only to a handful of new protein identifications.

Instead, we relied on the ADE strategy to break the saturation trend. We computed several physicochemical parameters for all distinct *B. henselae* proteins (see Supplementary

239 Methods). The statistical comparison of the parameters of 946 proteins identified by
240 OGEprot in the pilot phase versus the RNA-seq endpoint estimate of 1,353 expressed
241 proteins in both conditions provided evidence for a significant under-representation of short,
242 low abundant, basic and hydrophobic proteins. These areas of the proteome were
243 subsequently targeted by specific experimental approaches (see Supplementary Methods).
244 Under-representation with respect to length was targeted using size exclusion
245 chromatography (gelfiltration) (Brunner et al. 2007). These experiments added 83 new
246 protein identifications compared to the OGEprot pilot phase (Figure 3A, blue color). The
247 enrichment for shorter proteins can be appreciated in the upper left panel of Figure 3B. Low
248 abundant proteins were targeted using ProteoMiner (Fonslow et al. 2011; Guerrier et al.
249 2008). These experiments (Figure 3A, grey) helped to identify 42 additional proteins, which
250 were preferentially lower-abundant proteins as evidenced from the density distribution of
251 their Codon Adaptation Index (CAI) values (Sharp and Li 1987) (upper right panel in Figure
252 3B). Basic and membrane localized proteins were targeted using OFFGEL electrophoresis
253 at the peptide level (OGEpep). The 285 proteins newly added by the OGEpep experiments
254 (Figure 3A, red) were highly enriched for basic proteins (Figure 3B lower left panel) and
255 membrane proteins (with a high grand-average hydropathicity (gravy) value (Figure 3B lower
256 right panel).
257 Overall, we identified 1,250 distinct proteins requiring at least two PSMs per protein (Figure
258 S3), and only considering peptides that unambiguoulsy identify one bacterial protein (Qeli
259 and Ahrens 2010) (Table 1), i.e. 85% of the 1,467 distinct protein sequences. The FDRs at
260 the PSM, peptide and protein level are below 0.01%, 0.1% and 1%, respectively (Table 1).
261 Only few among the 1,228 proteins identified in the uninduced and 1,231 in the induced
262 condition were selectively expressed (Figure S4); these included several members of the
263 VirB/D4 T4SS in the induced condition. Compared to the expressed transcriptome, the
264 proteome coverage reaches 90% for both the uninduced and uninduced condition.

Although each experimental and computational approach contributed unique protein identifications to the final dataset (see Figure S5), for similar studies aiming to maximize coverage of an expressed proteome with a minimum number of experiments we recommend to use subcellular fractionation (Cyt and TM), perform OGEpep and measure each fraction twice using the exclusion list approach. This approach would identify 1,153 proteins, i.e. 92%, while requiring only 15% of the mass spectrometry runs needed to identify all 1,250 proteins.

Evidence for having reached an expressed proteome endpoint

Several lines of evidence indicated that the 1,250 distinct protein groups are very close to the complete proteome endpoint that is actively expressed under the investigated conditions. First, a comparison of the total number of PSM identifications showed that MS-GF+ added 67% more PSMs than Mascot-Percolator (Figure S3A). Yet, at the level of distinct peptides, this increase was smaller (+37%, Figure S3B), and amounted to a mere 3% or 33 additional proteins at the protein level (Figure S3C) despite having added several hundred thousand additional PSMs. Using a third search engine, Sequest, would have only added one additional protein for all experimental spectra. This indicates that, similar to the transcriptome, we have also measured the expressed proteome to saturation. The exponential model fitted to the eight OGEpep experiments (Figure 3A), supports this: doubling the number of PSMs on OGEpep samples (roughly 305,000 additional PSMs, i.e. approx. 36% more PSMs overall) would only identify 5 new proteins (red number on top of red dashed line, Figure 3A).

Second, our expressed proteome encompassed all proteins identified in three previous *B. henselae* proteomics studies (Eberhardt et al. 2009; Li et al. 2011; Rhomberg et al. 2004), while adding many more low abundant proteins (Figure S6A-C).

290 Third and most importantly, a comparison of the protein parameter distributions of the
291 datasets expressed protein-coding genes (1,353) and final expressed proteome (1,250)
292 showed that there is virtually no under-representation any more in those areas of the
293 proteome that we had specifically targeted, i.e. ADE successfully eliminated these
294 differences present in the OGEprot pilot study (Figure S7). Two examples to illustrate this
295 point: 1) for the parameter isoelectric point (pI) basic proteins are under-represented in the
296 OGEprot dataset. After carrying out the ADE approach, there is only a small difference
297 between the densities of the datasets final and expressed (Figure S7, top panels). 2) For the
298 parameter gravy, membrane proteins with one or more predicted transmembrane domains
299 (gravy values above 0.5) are under-represented in the OGEprot dataset. Again, after the
300 ADE approach, the densities for the datasets “expressed” and “final” are virtually identical
301 (Figure S7, middle panels). This comparison also showed that ADE could add proteins
302 encoded by genes that are expressed at lower levels under the conditions studied (Figure
303 S7, last panels). Two-dimensional density plots of the gene expression level versus the
304 parameters length, pI and gravy (Figure S8) for the dataset final expressed proteome
305 (1,250) versus not seen proteins (217) showed that there is still a noticeable tendency for
306 short and basic proteins to be enriched among genes with expression levels close to the
307 threshold whose proteins were not identified (Figure S8 A,B). These are not expected to be
308 detectable with the shotgun proteomics approach since short and basic proteins have fewer
309 tryptic peptides in the detectable range of the mass spectrometer. In contrast, for the two-
310 dimensional density plot with the protein parameter gravy (values above 0.5 are found in
311 proteins with transmembrane domains), we observed no bias (Figure S8C), indicative of a
312 complete membrane proteome coverage.
313 To correlate the gene expression level with the proteome coverage, we binned the protein-
314 coding genes according to gene expression strength (RPKM values) and plotted for each bin
315 the respective percentage of proteins identified (Figure 4). A clear correlation between

316 higher levels of gene expression with a higher success rate of protein identification can be
317 observed. However, several proteins of highly expressed genes were not identified: among
318 26 such cases from the five top expression bins, 23 had no conserved ortholog in
319 bartonellae, and 16 were located in a novel, plastic genome region (see next section).
320
321 **Integration of genome structure information and evolutionary conservation**
322 We projected transcriptomic and proteomic evidence, ortholog predictions and repeat
323 regions onto the *B. henselae* genome sequence (Figure 5), which contains a large prophage
324 region and three major genomic islands (Alsmark et al. 2004). Genes in such genomic
325 regions are often subject to regulation and become actively expressed only under specific
326 conditions (Juhas et al. 2009). Intriguingly, for 109 of the 198 genes that are located in these
327 four genomic regions we could not detect any expressed proteins (Figure 5, fourth ring). This
328 is a significant enrichment, given that only 227 annotated protein-coding genes did not
329 express any protein (p-value $<10^{-9}$, see Figure 5).
330 We next investigated whether the products of evolutionary conserved protein-coding genes
331 were enriched or selected against. In a comparison with *B. tribocorum*, *B. quintana* and *B.*
332 *grahamii*, 1,093 of the 1,488 *B. henselae* protein-coding genes were predicted to have an
333 ortholog (Engel et al. 2011), while 395 were not (Figure 5, third ring, turquoise bars). We
334 detected significant over-representation of genes lacking an ortholog (187 of 395) among the
335 227 protein-coding genes whose proteins were not identified (p-value $<10^{-9}$, Figure 5).
336 To extend the evolutionary conservation analysis beyond members of the genus *Bartonella*,
337 we relied on the eggNOG resource, which contains orthology information from 1,133
338 organisms including *B. henselae* (Powell et al. 2012). Among the 1,488 *B. henselae* proteins
339 only 55 proteins lack any functional annotation; they are a subset of the 395 without ortholog
340 (black bars, third ring, Figure 5). Strikingly, 52 of these 55 were not detected, again a
341 significant enrichment (p-value $<10^{-9}$). A significant number of the genes (16) encoding these

342 55 proteins clustered in a region from 1,612-1,674kbp that harbors 59 predicted ORFs (p-
343 value $<10^{-9}$) (yellow box, Figure 5). Location in this plastic, repeat-rich genome region
344 (orange bars, fourth ring) may lead to strong transcription of genes that do not represent a
345 bona fide protein-coding ORF.
346 The evolutionary conservation information provided by eggNOG together with high quality
347 experimental proteomics data, represents a particular useful combination to identify
348 candidates for over-predicted protein-coding genes in genome annotations: the densities of
349 the protein length distribution of the proteins not identified (217) was clearly separated from
350 that of the proteins seen (1,250) (Figure S9A). Among the proteins not seen, those that lack
351 any functional annotation are considerably shorter than those with a functional annotation
352 (Figure S9B). Since we can detect short proteins with our set-up (see density of the 150
353 shortest proteins detected compared to all, Figure S9C), the proteins that lack an ortholog
354 and any functional annotation may either only be expressed under different conditions, or
355 are potential over-predicted ORFs.
356
357 **Coverage of the membrane proteome and the VirB/D4 T4SS**
358 The membrane proteome serves many essential roles in cellular communication, transport,
359 adhesion to host cells and evasion of the host immune system. While accounting for up to
360 one third of the gene products, more than 50% of the druggable targets fall into this category
361 (Hopkins and Groom 2002). However, due to the amphipathic nature and low abundance of
362 membrane proteins, they are notoriously under-represented in proteomics studies (Helbig et
363 al. 2010; Poetsch and Wolters 2008; Tan et al. 2008).
364 To reach a high protein sequence coverage for membrane proteins, we used a combination
365 of trypsin and chymotrypsin in all membrane samples and furthermore applied proteolytic
366 digestion in 60% (v/v) Methanol to improve cleavability of hydrophobic proteins (Fischer et
367 al. 2006) (Supplementary Methods). Among 924 proteins identified in the first four pilot

368 phase experiments (63% of all distinct proteins), 182 contained predicted transmembrane
369 domains (54%, Figure 6A, upper panel). However, the ADE approach was able to eliminate
370 this under-representation of membrane proteins: among the final 1,250 identified proteins
371 (85% of all distinct annotated proteins) 289 of the 338 distinct proteins with one or more
372 predicted transmembrane regions were found, i.e. 86% (Figure 6A, lower panel; Figure
373 S10A). Notably, the OGEpep fractionation regimen was particularly successful to identify
374 membrane proteins. We also identified 54 of the 58 predicted secreted proteins (95%).
375 These include many proteins for which PSORTb (Yu et al. 2010) predicts localization in the
376 membrane space, and where other studies could confirm their localization in inner or outer
377 membrane, periplasm or the extra-cellular space (Figure S10B). Together with the striking
378 result that transmembrane proteins with high gravy values are not over-represented among
379 the 217 non-identified proteins compared to 1,250 seen proteins (see Figure S8C), the data
380 suggested that we have identified a complete membrane proteome expressed under two
381 specific conditions.
382 This includes all eleven protein members encoded by the *virB*/*D4* operon in the induced
383 condition (Figure 6B). To our knowledge, this is the first complete coverage of this important
384 molecular machinery spanning both inner and outer membrane by a shotgun proteomics
385 approach. We also detected all seven *Bartonella* effector proteins (Beps), which are
386 secreted by the VirB/D4 T4SS into eukaryotic host cells (Figure 6B). In contrast, many
387 proteins of the Trw complex, a second *B. henselae* T4SS that is essential for the infection of
388 erythrocytes (Vayssier-Taussat et al. 2010) but dispensable under the conditions studied,
389 were not detected (9 of 24, 38%) (Figure 5, first and fifth ring), nor was their expression
390 regulated (Figure S11).
391 When we assessed the level of induction at the RNA and protein level, we observed that the
392 induction of *virB*/*D4* and *bep* operons, which are direct targets of the transcriptional regulator
393 BatR, seemed to be more prominent at the protein level. They also included more cases with

394 statistical significance of the up-regulation (Figure 6B, log_2 fold changes, left panel). A
395 comparison of the log_2 fold changes at the RNA level versus those at the protein level
396 indicated that several of the *virB*/*D4* and *bep* genes appear to be regulated preferentially at
397 the post-transcriptional level, indicated in Figure 6C by their position close to the vertical
398 axis.
399
400 The ability to identify complete membrane proteomes of prokaryotes has important
401 implications for studying their expression under different conditions in a quantitative fashion.
402 Ideally, such a task would be performed with the more sensitive targeted proteomics
403 approach (Schmidt et al. 2011), which typically relies on predicted PTPs using tools like
404 PeptideSieve (Mallick et al. 2007). Our data indicate that a comprehensive discovery
405 proteomics approach adds clear value with respect to experimentally identified PTPs as we
406 could identify peptides for 145 proteins for which PeptideSieve predicted no PTP (see
407 Supplementary Methods). We provide the proteomics and transcriptomics data with results
408 of several prediction algorithms (Table S5A), and all experimentally identified peptides
409 (Table S5B), from which the best-suited PTPs can be selected using available guidelines
410 (Picotti and Aebersold 2012).
411
412 **Identification of differentially expressed proteins**
413 Our in depth proteome analysis precluded the measurement of biological replicates. We thus
414 relied on DESeq to identify the most significantly differentially regulated proteins between
415 induced and uninduced states (see Methods). The top 10% differentially expressed proteins
416 (Table S6), including 68 up-regulated (red dots), and 57 down-regulated proteins (green
417 dots) in the induced condition, are highlighted in Figure 7.
418 Among these 125, 36 transmembrane and 12 secreted proteins were found, a significant
419 enrichment (p-value <0.0018) compared to 343 membrane and secreted proteins among the

1,250 proteins. A striking feature was the strong regulation of different families of autotransporters, which rely on the type V secretion pathway for their delivery to the surface of Gram-negative bacteria (Leyton et al. 2012). These included two representatives of the trimeric autotransporter adhesins (BH01490, BH01510), a class of virulence factors essential for *Bartonella* pathogenicity (Franz and Kempf 2011). Furthermore, 7 of 10 proteins with an autotransporter beta domain (as predicted by SMART version 7, (Letunic et al. 2012)) were among the top 10% differentially regulated proteins (6 up-regulated, 1 down-regulated; yellow dots, Figure 7; Table S6), i.e. a significant enrichment (p-value $<4x10^{-7}$). BH13020, BH13180 and BH13010 were the top three up-regulated proteins, which ranked even higher than members of the VirB/D4 operon. While less is known about the role of this family of autotransporters in *Bartonella*, they were found to be up-regulated during infection of endothelial cells (Quebatte et al. 2010) and may be involved in adhesion to host cells (Litwin et al. 2007). Finally, two of the four outer membrane proteins of the hemin binding protein family (HbpC and HbpB) were found. HbpC was shown to protect *B. henselae* against hemin toxicity and to play a role during host infection (Roden et al. 2012).

The top 10% regulated proteins included 6 of the 7 Beps and all VirB/D4 T4SS proteins except VirB3. For this small protein (103 amino acids) with one predicted transmembrane domain we only found 4 spectra, all in the induced condition. This indicates that a large experimental effort is required to detect proteins that combine several parameters which complicate their mass spectrometric identification with shotgun proteomics, i.e. they are short, basic and hydrophobic. Another protein exclusively identified in the induced condition is BH13250, a hypothetical protein with a transmembrane domain (Table S6). Its location right upstream of the *virB/D4* operon is conserved in other *Bartonella*, suggesting that it may potentially carry out a yet to be determined function as virulence factor. Finally, another interesting upregulated protein is RpoH1 (BH15210), an alternative RNA-polymerase sigma factor 32. It has recently been identified as an essential component for the expression of the

446 VirB/D4 T4SS by independent approaches (M. Quebatte et al., personal communication). A
447 role in virulence has been documented for its gene in an *in vivo* mouse infection model for
448 the closely related *Brucella* (Delory et al. 2006).
449

CONCLUSION

Using a discovery proteomics approach, the expressed proteome of *B. henselae* was exhaustively studied under two conditions that mimic those encountered in different hosts. The saturated transcriptome analysis of RNA extracted from matched samples provided the best-possible endpoint estimate for the number of actively transcribed protein-coding genes. ADE was able to virtually eliminate the biases of commonly under-represented short, basic, and particularly lower abundant and membrane protein classes, all of which are experimentally tractable. Based on a very stringent FDR at the PSM level, we identified 85% of all distinct, annotated proteins and about 90% compared to the expressed protein-coding genes in the two conditions. Several lines of evidence indicated that this is very close to all proteins that can be identified by a discovery proteomics approach with current technology. This is best illustrated by the complete membrane proteome coverage, including evidence for all members of the important VirB/D4 T4SS. The analysis of the genome organization revealed that genes whose transcripts were detected, but not their corresponding protein products, were highly enriched in genomic islands. Information regarding evolutionary conservation provided evidence for preferential expression of genes with a predicted ortholog. In contrast, genes that lacked an ortholog and functional annotation were mostly not observed at the protein level, suggesting possible over-prediction in genome annotations.

Our report is the second complete expressed proteome reported (de Godoy et al. 2008). Using a similarly extensive fractionation strategy, our matched transcriptomics and proteomics data correlated quite well (r=0.57) while identifying the VirB/D4 T4SS as a prominent target of post-transcriptional regulation. The rigorous approach to sequence transcriptome and proteome to saturation and to provide proof for having eliminated observed biases at the protein level is unique. It supports a recent perspective article showing that up to 90% of an expressed proteome ("nearly complete") can be measured

476 quite fast (Mann et al. 2013), but the remaining 10% require extensive effort. It also
477 underlines that the difference between “comprehensive” and complete can be quite large, in
478 particular with respect to coverage of the membrane proteome (Beck et al. 2011). The
479 higher coverage of distinct annotated proteins (85%) compared to the proteome expressed
480 by haploid and diploid yeast (67%) suggests that prokaryotes express a higher fraction of
481 the encoded proteins potentially reflecting their need to quickly adapt to changing conditions.
482 This fraction may be lower for more complex prokaryotes.
483 The data attest to the value of a discovery proteomics approach in providing experimentally
484 identified PTPs beyond those predicted *in silico*. The sensitive quantitative measurement of
485 such PTPs by SRM holds particular promise to be able to screen entire bacterial
486 surfaceomes and to identify targets for novel anti-infectives. Ideally, such studies would be
487 carried out using *in vivo* infection models. Enabled by the consideration of organism-specific
488 peptide information (Delmotte et al. 2010), they will bring the analysis of mixed *in vivo*
489 proteomes within reach and complement the power of dual RNA-seq (Westermann et al.
490 2012) for this task. We expect that the strategy described here will be useful for some of
491 these exciting applications.
492

METHODS

Bacterial growth and subcellular fractionation

B. henselae strain MQB307 harbors a deletion of the response regulator *batR* (BH00620) and its cognate sensor histidine kinase *batS* (BH00610), and carries a plasmid-encoded copy of *batR* under the control of an IPTG inducible promoter (for details, see Supplementary Methods, Tables S1, S2). MQB307 was grown on Columbia blood agar (CBA) plates supplemented with 30 mg/l kanamycin with (induced condition) or without (uninduced condition) 500 μM IPTG at 35°C and 5% CO_2 for 60 h. The subcellular fractionation was performed as previously described (Rhomberg et al. 2004) (Supplementary Methods). To maximize the recovery of membrane proteins, the total membrane fraction (TM) was further separated into inner membrane (IM) and outer membrane (OM) fraction.

RNA extraction and whole transcriptome sequencing

RNA was isolated from bacterial cells as described (Quebatte et al. 2010). Whole Transcriptome libraries were produced using the Ribominus kit, Bacteria Module (Invitrogen),and the SOLiD™ Total RNA-seq kit (Applied Biosystems). Briefly, cDNA libraries were size selected and amplified for 18 cycles of PCR. The whole transcriptome library was used for emulsion-PCR based on a concentration of 0.5 pM. Sequencing beads were pooled and loaded on a full SOLiD™-4 slide; between 55-87 million 50 base sequencing reads were generated per library (Table S3). For details, see Supplementary Methods.

RNA-seq data processing and transcriptome coverage analysis

The sequenced reads were mapped to the genome sequence of *B. henselae* Houston-1 strain using the BioScope 1.3.1 mapping pipeline. Among all uniquely mapping reads, those of lower quality were removed (for more detail, see Supplementary Methods, Figure S12). The count data summary for annotated *B. henselae* ORFs was generated using the HTSeq package. To create Figure 2A, the filtered reads were shuffled and sequentially mapped to the genome; a protein-coding ORF was classified as expressed when accumulating five or more distinct reads in the 5' end of the ORF. Based on this data, nonlinear regression models were constructed to estimate the effect of doubling the number of reads. For details, see Supplementary Methods.

Protein and peptide fractionation & mass spectrometry

The subcellular fractions ($Cyt_{u/i}$, $TM_{u/i}$, $IM_{u/i}$, $OM_{u/i}$) were further fractionated biochemically, including OFFGEL electrophoresis at the protein (OGEprot) and peptide level (OGEpep), and size exclusion chromatography (SEC, "gelfiltration"). To enrich for low-abundant proteins, we used the ProteoMiner approach (Guerrier et al. 2008). More detail on the biochemical fractionations, digest conditions and the mass spectrometry set-up is given in the Supplementary Methods and Figure S13. Samples were injected into an Eksigent-nano-HPLC system (Eksigent Technologies) by an autosampler, separated on a self-made reverse-phase tip column packed with C18 material, and acquired on an LTQ-Orbitrap XL or LTQ-ICR-FT-Ultra mass spectrometer (both Thermo Scientific).

Database searching and data processing

To minimize the chance for false positive assignments spectra were searched against a combined database (1,488 *B. henselae* proteins, 3,336 sheep proteins, a positive control (myc-gfp), and sequences of 256 common contaminants (keratins, trypsin, etc.)) either with Mascot (version 2.3.0, Matrix Science) or with MS-GF+ (MS-GFDB v7747). For Mascot, data were further post-processed with Percolator (Brosch et al. 2009). Based on the target-decoy search approach, a Percolator/MS-GF+ score cutoff was determined that resulted in an estimated 0.01% FDR at the PSM level. All PSMs above this cutoff were classified with the PeptideClassifier software (Qeli and Ahrens 2010), and only peptides (tryptic or semi-tryptic) that unambiguously imply one bacterial protein sequence were considered (Table 1). For details, see Supplementary Methods.

ADE analysis

Exponential curves were fitted to each block of experiments with a shared biochemical fractionation regimen to find a saturation threshold (Figure 3A). We then used this fit to predict the saturation beyond the point of experimentally observed PSMs for each biochemical fractionation regimen (Figure 3A, dashed lines). For details on the exponential model, approximating confidence bands, density estimation of physicochemical parameters, and computation of physicochemical parameters and other protein sequence features, see Supplementary Methods.

Statistical analysis

Statistical tests were performed using the statistical software R 2.15.2 (www.R-project.org). All reported p-values are from hypergeometric tests and are adjusted for multiple testing

561 controlling the corresponding FDR (Benjamini and Hochberg 1995). Significance is based on
562 an alpha level of 5%.
563
564 **Transcript and protein abundance estimation**
565 Transcript abundance was estimated via RPKM values calculated similar to (Mortazavi et al.
566 2008). The sum of mapped and filtered reads per gene was divided by its length (in
567 kilobases) and the sum of reads for all *B. henselae* protein-coding genes (in million reads).
568 Relative protein abundance (in ppm, see Figure S6C) was estimated based on spectral
569 counts as described (Schrimpf et al. 2009).
570
571 **Orthologs, sequence repeats, and functional protein classification**
572 Orthologous genes conserved in *B. henselae*, *B. tribocorum,* and *B. grahamii* were taken
573 from (Engel et al. 2011). To find duplicated regions of 50nt or longer in the *B. henselae*
574 genome, we used RepSeek (version 6.5, (Achaz et al. 2007)). For functional protein
575 classification, we relied on the eggNOG resource (http://eggnog.embl.de). For details, see
576 Supplementary Methods.
577
578 **Differential expression analysis**
579 Differential transcript and protein expression analysis was carried out with the R package
580 DESeq (version 1.6.1, (Anders and Huber 2010)). Our description of condition-specific
581 complete expressed proteomes precluded the analysis of biological replicates. Since DESeq
582 ranks proteins according to statistical significance, i.e. the top-ranked proteins are observed
583 by many spectra, we minimized the potential to erroneously identify differentially expressed
584 proteins by chance. On the other hand, without replicates we lack the power to detect lower
585 expressed, truly differentially regulated proteins.
586
587 **DATA ACCESS**
588 RNA-seq data is accessible under the GEO Series accession number GSE44564.
589 Proteomics data associated with this manuscript can be downloaded from
590 ProteomeXchange under accession PXD000153.
591

ACKNOWLEDGEMENTS

We thank Dr. Ermir Qeli for initial work on the project, Dr. Sangtae Kim (PNNL, USA) for the MS-GF+ software, and Drs. Bernd Wollscheid (IMSB, ETH Zürich), Aurelien Carlier and Gabriella Pessi (Institute of Plant Biology, UZH) for critical reading of the manuscript. CHA acknowledges support from the Swiss National Science Foundation (SNF) under grant 31003A_130723. A part of the research was performed using EMSL, a national scientific user facility sponsored by the Department of Energy's Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory.

FIGURE AND TABLE LEGENDS

FIGURES

Figure 1: Overview of the complete expressed proteome discovery workflow.
A. Extraction of RNA and proteins from matched samples, transcriptome analysis. Total RNA and proteins were extracted in parallel from bacteria grown either under uninduced or induced conditions (schematically shown by black knobs representing the VirB/D4 T4SS). Protein extracts were sub-fractionated into cytoplasmic (Cyt), total membrane (TM), inner (IM) and outer membrane (OM) fractions. To estimate an upper bound for the number of actively transcribed protein-coding genes, the transcriptome was sequenced to saturation using RNA-seq.
B. Analysis-driven experimentation (ADE). In a first pilot phase, samples are analyzed by LC-MS/MS. Under-represented proteome areas are identified based on a statistical analysis comparing experimentally identified proteins to all expressed proteins (the estimated RNA-seq endpoint indicated by the orange dashed line within an error envelope). All distinct annotated proteins are indicated by the black, dashed line. Subsequently, these areas are investigated by targeted experiments, aiming to overcome the saturation trend.
C. Integrative data analysis. Data from the expressed proteome are integrated with genomic, transcriptomics, orthology, and other information to enable further analyses.

Figure 2: Transcriptome coverage by RNA-seq.
A. Saturated coverage of protein-coding genes. An estimate of the number of actively expressed protein-coding genes based on the number of uniquely mapped RNA-seq reads is shown for both conditions and biological replicates.
B. Density distribution of RPKM values. In addition, boxplots representing the expression level of the eleven members of the *virB/D4* operon are shown for the uninduced (black), and induced (red) condition. For clarity, we only show data for the sample pair uninduced2/induced2.

Figure 3: Overcoming the saturation of protein identifications using ADE-guided shotgun proteomics.
A. Increase of distinct identified proteins given the number of PSMs observed in different experiments. We fitted an exponential curve (see Methods) to all experiments for a given biochemical fractionation in order to find a saturation limit (see colored numbers on the right-

636 hand side). We also approximated confidence bands for the fitted points (thin lines, see
637 Methods). The black dashed line at the top signifies the total number of distinct *B. henselae*
638 proteins (1,467), the orange dashed line below represents the estimated RNA-seq endpoint
639 of expressed distinct proteins (1,353) including a ±2.5% error envelope (orange shaded
640 area).
641 B. Density estimates of four physicochemical protein parameters for different protein
642 subsets. The parameter density for proteins newly identified by the ADE approach is
643 contrasted to that of all expressed proteins (orange) and those identified in pilot experiments
644 using OGEprot (green). The most important aspects of over- or under-representation can be
645 seen on the abscissa; they indicate that the targeted experiments successfully add new
646 protein identifications in areas of the proteome that were under-represented in the pilot
647 experiments. For details on the density estimation and the bootstrap confidence bands
648 (shaded areas) see Supplementary Methods.
649
650 **Figure 4:** Correlation of gene expression strength and successful protein identification rate.
651 Protein-coding genes are binned according to strength of gene expression (the maximum
652 RPKM value of both states). The success rate in identifying the encoded proteins in each bin
653 is represented by the blue area of the bars; orange dots above the barplot indicate the
654 respective percentage. The numbers above the bars show how many proteins were not
655 identified within a given bin (for a total of 217 distinct proteins).
656
657 **Figure 5:** Integration of expression evidence with structural genome information and
658 evolutionary conservation.
659 Genes whose proteins were not identified cluster in specific regions of the *B. henselae*
660 genome. Outer ring: genes whose proteins were identified (light blue), or not identified (red).
661 Second ring: protein-coding genes classified by the RNA-seq analysis as expressed (grey),
662 or not (dark green). Third ring: genes without a detectable ortholog among species of
663 lineage 4 of the genus *Bartonella* (Engel et al. 2011) (turquoise), and genes without any
664 functional annotation by the eggNOG classification (black). Fourth ring: repeat regions
665 identified by RepSeek (orange) (Vallenet et al. 2006), and rRNA repeat regions (light
666 orange). Fifth ring: location of a prophage region (ochre), three genomic islands (blue), the
667 *virB/D4* and *trw* operons (skyblue), and a novel genomic region enriched in repeats as well
668 as highly expressed genes whose encoded proteins were not identified (yellow). The results
669 of hypergeometric tests for selected datasets are also shown (asterisks indicate statistically
670 significant enrichment, see text). For the hypergeometric test, we used all possible protein-

671 coding genes for the identified “seen” proteins (1,250 distinct proteins encoded by 1,261
672 gene models) and “not seen” proteins (217 distinct proteins encoded by 227 gene models).
673 The circular plot was generated using DNAPlotter (Carver et al. 2009).
674
675 **Figure 6:** Membrane proteome coverage and dynamics.
676 A. Comparison of the membrane proteome coverage achieved in four pilot experiments
677 (upper panel) and the final dataset (lower panel). Membrane proteins are binned according
678 to the number of predicted transmembrane domains; the percentage of proteins identified
679 per bin is shown above each bar. The legends summarize the respective coverage achieved
680 comparing the respective dataset (pilot phase/final) against all distinct proteins and for the
681 subset of proteins with transmembrane domains. Membrane proteins are under-represented
682 in the pilot phase but not in the final dataset.
683 B. Transcript and protein expression changes of the *virB/D4* T4SS and downstream *bep*
684 operon. Operon structures (upper panel) are drawn to scale. The lower left panel shows the
685 $\log_2$ fold changes at the transcript and protein level for the induced versus uninduced state
686 (the ∞ indicates that the protein was only identified in the induced condition). Fold changes
687 and significance were calculated with DESeq. Regulation at the protein level appears to be
688 more pronounced compared to the transcript level. The lower right panel visualizes the
689 protein expression changes upon induction onto a schematic representation of the
690 assembled VirB/D4 T4SS using different shades of blue.
691 C. Comparison of expression changes at transcript and protein level. The respective $\log_2$
692 fold changes based on the RPKM values and normalized spectral counts are shown.
693 Members of the VirB/D4 T4SS are shown in blue (BH13360 in light blue), *Bartonella* effector
694 proteins (Beps) in dark blue. Three proteins that exhibited the most significant differential
695 expression (Table S6, Figure 7) are also shown with their identifiers.
696
697 **Figure 7:** Differential protein expression analysis.
698 The $\log_2$ fold change of the expression of all experimentally identified *B. henselae* proteins in
699 the induced versus uninduced condition is shown against the mean normalized spectral
700 count (MA plot). The 10% most significant differentially expressed proteins are highlighted,
701 including 68 up-regulated proteins (red dots) and 57 down-regulated proteins (green dots).
702 Selected regulated proteins are highlighted in different colors: members of the VirB/D4 T4SS
703 (blue dots), Beps (dark blue dots), and several proteins containing autotransporter beta-
704 domains (yellow dots). Proteins in these categories that rank below the 10% cutoff are
705 shown as open circles.

706 **TABLES**

707

708 **Table 1:** Summary of identified PSMs, peptides and proteins and estimated FDR levels

709

	No. of PSMs	No. of distinct peptides	No. of distinct proteins*
Class 1a	747,352	43,193	1,240
Class 3a	7,356	283	10
Class 3b	12,161	663	n.a.
Total *B. henselae*	766,869	44,139	1,250
Decoy hits	54	42	7
estimated FDR	**< 0.01%**	**< 0.1%**	**< 1.0%**

710

711 The total number of PSMs, distinct peptides and distinct proteins is shown, further separated by
712 peptide evidence class (Grobei et al. 2009). We only considered proteins implied by class 1a and 3a
713 peptides, not those implied by ambiguous class 3b peptides (n.a.). *Protein groups identified by 3a
714 peptides are unique protein sequences that can be encoded by two or more distinct gene models.
715 The 1,250 experimentally identified proteins are encoded by 1,261 gene models; the 217 non-
716 identified proteins are encoded by 227 gene models (in total: 1,467 distinct proteins are encoded by
717 1,488 protein-coding genes).

718

REFERENCES

Achaz, G., Boyer, F., Rocha, E.P., Viari, A., and Coissac, E. 2007. Repseek, a tool to retrieve approximate repeats from large DNA sequences. *Bioinformatics* 23: 119-121.

Ahrens, C.H., Brunner, E., Qeli, E., Basler, K., and Aebersold, R. 2010. Generating and navigating proteome maps using mass spectrometry. *Nat Rev Mol Cell Biol* 11: 789-801.

Alsmark, C.M., Frank, A.C., Karlberg, E.O., Legault, B.A., Ardell, D.H., Canback, B., Eriksson, A.S., Naslund, A.K., Handley, S.A., Huvet, M. et al. 2004. The louse-borne human pathogen *Bartonella quintana* is a genomic derivative of the zoonotic agent *Bartonella henselae*. *Proc Natl Acad Sci U S A* 101: 9716-9721.

Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biol* 11: R106.

Balgley, B.M., Laudeman, T., Yang, L., Song, T., and Lee, C.S. 2007. Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy. *Mol Cell Proteomics* 6: 1599-1608.

Becher, D., Hempel, K., Sievers, S., Zuhlke, D., Pane-Farre, J., Otto, A., Fuchs, S., Albrecht, D., Bernhardt, J., Engelmann, S. et al. 2009. A proteomic view of an important human pathogen--towards the quantification of the entire *Staphylococcus aureus* proteome. *PLoS One* 4: e8176.

Beck, M., Claassen, M., and Aebersold, R. 2011. Comprehensive proteomics. *Curr Opin Biotechnol* 22: 3-8.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. . *J Roy Statist Soc Ser B* 57: 289-300.

Brosch, M., Yu, L., Hubbard, T., and Choudhary, J. 2009. Accurate and sensitive peptide identification with Mascot Percolator. *J Proteome Res* 8: 3176-3181.

Brunner, E., Ahrens, C.H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., Deutsch, E.W., Panse, C., de Lichtenberg, U., Rinner, O. et al. 2007. A high-quality catalog of the *Drosophila melanogaster* proteome. *Nat Biotechnol* 25: 576-583.

Carver, T., Thomson, N., Bleasby, A., Berriman, M., and Parkhill, J. 2009. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 25: 119-120.

de Godoy, L.M., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Frohlich, F., Walther, T.C., and Mann, M. 2008. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455: 1251-1254.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., and Vogel, C. 2009. Global signatures of protein and mRNA expression levels. *Mol Biosyst* 5: 1512-1526.

Delmotte, N., Ahrens, C.H., Knief, C., Qeli, E., Koch, M., Fischer, H.M., Vorholt, J.A., Hennecke, H., and Pessi, G. 2010. An integrated proteomics and transcriptomics reference dataset provides new insights into the *Bradyrhizobium japonicum* bacteroid metabolism in soybean root nodules. *Proteomics* 10: 1391-1400.

Delory, M., Hallez, R., Letesson, J.J., and De Bolle, X. 2006. An RpoH-like heat shock sigma factor is involved in stress response and virulence in *Brucella melitensis* 16M. *J Bacteriol* 188: 7707-7710.

Eberhardt, C., Engelmann, S., Kusch, H., Albrecht, D., Hecker, M., Autenrieth, I.B., and Kempf, V.A. 2009. Proteomic analysis of the bacterial pathogen *Bartonella henselae* and identification of immunogenic proteins for serodiagnosis. *Proteomics* 9: 1967-1981.

Engel, P., Salzburger, W., Liesch, M., Chang, C.C., Maruyama, S., Lanz, C., Calteau, A., Lajus, A., Medigue, C., Schuster, S.C. et al. 2011. Parallel evolution of a type IV secretion system in radiating lineages of the host-restricted bacterial pathogen *Bartonella*. *PLoS Genet* 7: e1001296.

Fischer, F., Wolters, D., Rogner, M., and Poetsch, A. 2006. Toward the complete membrane proteome: high coverage of integral membrane proteins through transmembrane peptide detection. *Mol Cell Proteomics* 5: 444-453.

Fonslow, B.R., Carvalho, P.C., Academia, K., Freeby, S., Xu, T., Nakorchevsky, A., Paulus, A., and Yates, J.R., 3rd. 2011. Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *J Proteome Res* 10: 3690-3700.

Franz, B. and Kempf, V.A. 2011. Adhesion and host cell modulation: critical pathogenicity determinants of *Bartonella henselae*. *Parasit Vectors* 4: 54.

Giannone, R.J., Huber, H., Karpinets, T., Heimerl, T., Kuper, U., Rachel, R., Keller, M., Hettich, R.L., and Podar, M. 2011. Proteomic characterization of cellular and molecular processes that enable the *Nanoarchaeum equitans* -*Ignicoccus hospitalis* relationship. *PLoS ONE* 6: e22942.

Grobei, M.A., Qeli, E., Brunner, E., Rehrauer, H., Zhang, R., Roschitzki, B., Basler, K., Ahrens, C.H., and Grossniklaus, U. 2009. Deterministic protein inference for shotgun proteomics data provides new insights into *Arabidopsis* pollen development and function. *Genome Res* 19: 1786-1800.

Guerrier, L., Righetti, P.G., and Boschetti, E. 2008. Reduction of dynamic protein concentration range of biological extracts for the discovery of low-abundance proteins by means of hexapeptide ligand library. *Nat Protoc* 3: 883-890.

Harms, A. and Dehio, C. 2012. Intruders below the radar: molecular pathogenesis of *Bartonella* spp. *Clin Microbiol Rev* 25: 42-78.

Helbig, A.O., Heck, A.J., and Slijper, M. 2010. Exploring the membrane proteome--challenges and analytical strategies. *J Proteomics* 73: 868-878.

Hopkins, A.L. and Groom, C.R. 2002. The druggable genome. *Nat Rev Drug Discov* 1: 727-730.

Jaffe, J.D., Stange-Thomann, N., Smith, C., DeCaprio, D., Fisher, S., Butler, J., Calvo, S., Elkins, T., FitzGerald, M.G., Hafez, N. et al. 2004. The complete genome and proteome of *Mycoplasma mobile*. *Genome Res* 14: 1447-1461.

Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* 33: 376-393.

Kim, S., Gupta, N., and Pevzner, P.A. 2008. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res* 7: 3354-3363.

Kim, S., Mischerikow, N., Bandeira, N., Navarro, J.D., Wich, L., Mohammed, S., Heck, A.J., and Pevzner, P.A. 2010. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol Cell Proteomics* 9: 2840-2852.

Kristensen, D.B., Brond, J.C., Nielsen, P.A., Andersen, J.R., Sorensen, O.T., Jorgensen, V., Budin, K., Matthiesen, J., Veno, P., Jespersen, H.M. et al. 2004. Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol Cell Proteomics* 3: 1023-1038.

Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. 2005. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* 6: 577-583.

Letunic, I., Doerks, T., and Bork, P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 40: D302-305.

Leyton, D.L., Rossiter, A.E., and Henderson, I.R. 2012. From self sufficiency to dependence: mechanisms and factors important for autotransporter biogenesis. *Nat Rev Microbiol* 10: 213-225.

Li, D.M., Liu, Q.Y., Zhao, F., Hu, Y., Xiao, D., Gu, Y.X., Song, X.P., and Zhang, J.Z. 2011. Proteomic and bioinformatic analysis of outer membrane proteins of the protobacterium *Bartonella henselae* (Bartonellaceae). *Genet Mol Res* 10: 1789-1818.

Litwin, C.M., Rawlins, M.L., and Swenson, E.M. 2007. Characterization of an immunogenic outer membrane autotransporter protein, Arp, of *Bartonella henselae*. *Infect Immun* 75: 5255-5263.

Maier, T., Schmidt, A., Guell, M., Kuhner, S., Gavin, A.C., Aebersold, R., and Serrano, L. 2011. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol Syst Biol* 7: 511.

Mallick, P., Schirle, M., Chen, S.S., Flory, M.R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T. et al. 2007. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* 25: 125-131.

Malmstrom, J., Beck, M., Schmidt, A., Lange, V., Deutsch, E.W., and Aebersold, R. 2009. Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*. *Nature* 460: 762-765.

Mann, M., Kulak, N.A., Nagaraj, N., and Cox, J. 2013. The coming age of complete, accurate and ubiquitous proteomes. *Molecular Cell* 49: 583-590.

Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., and Bahler, J. 2012. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* 151: 671-683.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5: 621-628.

Nesvizhskii, A.I. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* 73: 2092-2123.

Norrby, S.R., Nord, C.E., and Finch, R. 2005. Lack of development of new antimicrobial drugs: a potential serious threat to public health. *Lancet Infect Dis* 5: 115-119.

Picotti, P. and Aebersold, R. 2012. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nat Methods* 9: 555-566.

Poetsch, A. and Wolters, D. 2008. Bacterial membrane proteomics. *Proteomics* 8: 4100-4122.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T. et al. 2012. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* 40: D284-289.

Qeli, E. and Ahrens, C.H. 2010. PeptideClassifier for protein inference and targeted quantitative proteomics. *Nat Biotechnol* 28: 647-650.

Quebatte, M., Dehio, M., Tropel, D., Basler, A., Toller, I., Raddatz, G., Engel, P., Huser, S., Schein, H., Lindroos, H.L. et al. 2010. The BatR/BatS two-component regulatory system controls the adaptive response of *Bartonella henselae* during human endothelial cell infection. *J Bacteriol* 192: 3352-3367.

Reiter, L., Claassen, M., Schrimpf, S.P., Jovanovic, M., Schmidt, A., Buhmann, J.M., Hengartner, M.O., and Aebersold, R. 2009. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* 8: 2405-2417.

Rhomberg, T.A., Karlberg, O., Mini, T., Zimny-Arndt, U., Wickenberg, U., Rottgen, M., Jungblut, P.R., Jeno, P., Andersson, S.G., and Dehio, C. 2004. Proteomic analysis of the sarcosine-insoluble outer membrane fraction of the bacterial pathogen *Bartonella henselae*. *Proteomics* 4: 3021-3033.

Roden, J.A., Wells, D.H., Chomel, B.B., Kasten, R.W., and Koehler, J.E. 2012. Hemin binding protein C is found in outer membrane vesicles and protects *Bartonella henselae* against toxic concentrations of hemin. *Infect Immun* 80: 929-942.

Savas, J.N., Stein, B.D., Wu, C.C., and Yates, J.R., 3rd. 2011. Mass spectrometry accelerates membrane protein analysis. *Trends Biochem Sci* 36: 388-396.

Schmidt, A., Beck, M., Malmstrom, J., Lam, H., Claassen, M., Campbell, D., and Aebersold, R. 2011. Absolute quantification of microbial proteomes at different states by directed mass spectrometry. *Mol Syst Biol* 7: 510.

Schrimpf, S.P., Weiss, M., Reiter, L., Ahrens, C.H., Jovanovic, M., Malmstrom, J., Brunner, E., Mohanty, S., Lercher, M.J., Hunziker, P.E. et al. 2009. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol* 7: e48.

Schulein, R. and Dehio, C. 2002. The VirB/VirD4 type IV secretion system of *Bartonella* is essential for establishing intraerythrocytic infection. *Mol Microbiol* 46: 1053-1067.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. 2011. Global quantification of mammalian gene expression control. *Nature* 473: 337-342.

Sharp, P.M. and Li, W.H. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15: 1281-1295.

Tan, S., Tan, H.T., and Chung, M.C. 2008. Membrane proteins and membrane proteomics. *Proteomics* 8: 3924-3932.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., and Medigue, C. 2006. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* 34: 53-65.

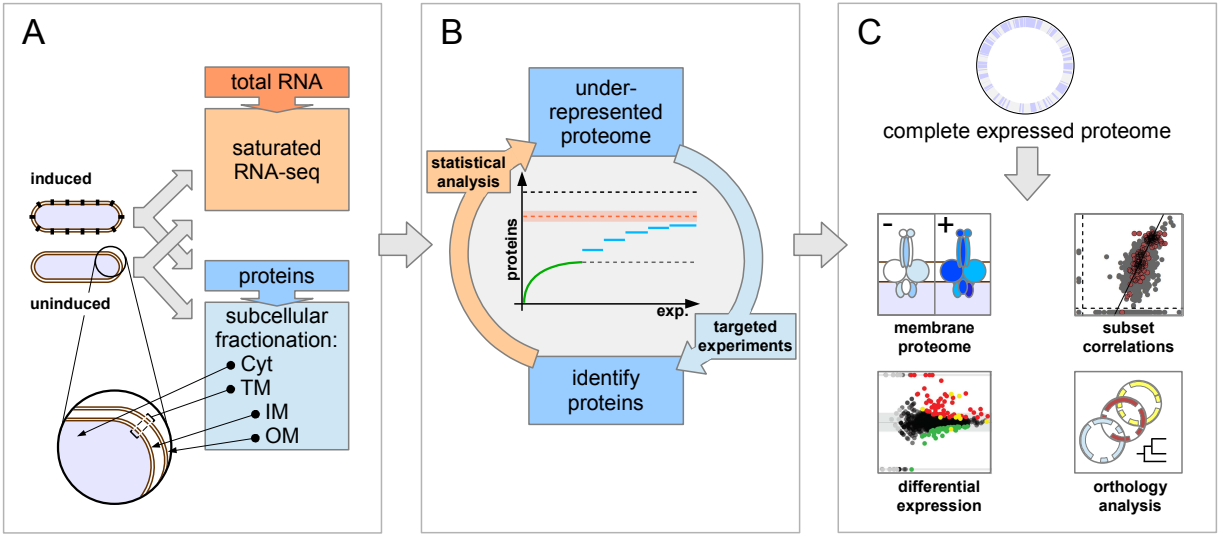
Vayssier-Taussat, M., Le Rhun, D., Deng, H.K., Biville, F., Cescau, S., Danchin, A., Marignac, G., Lenaour, E., Boulouis, H.J., Mavris, M. et al. 2010. The Trw type IV secretion system of *Bartonella* mediates host-specific adhesion to erythrocytes. *PLoS Pathog* 6: e1000946.

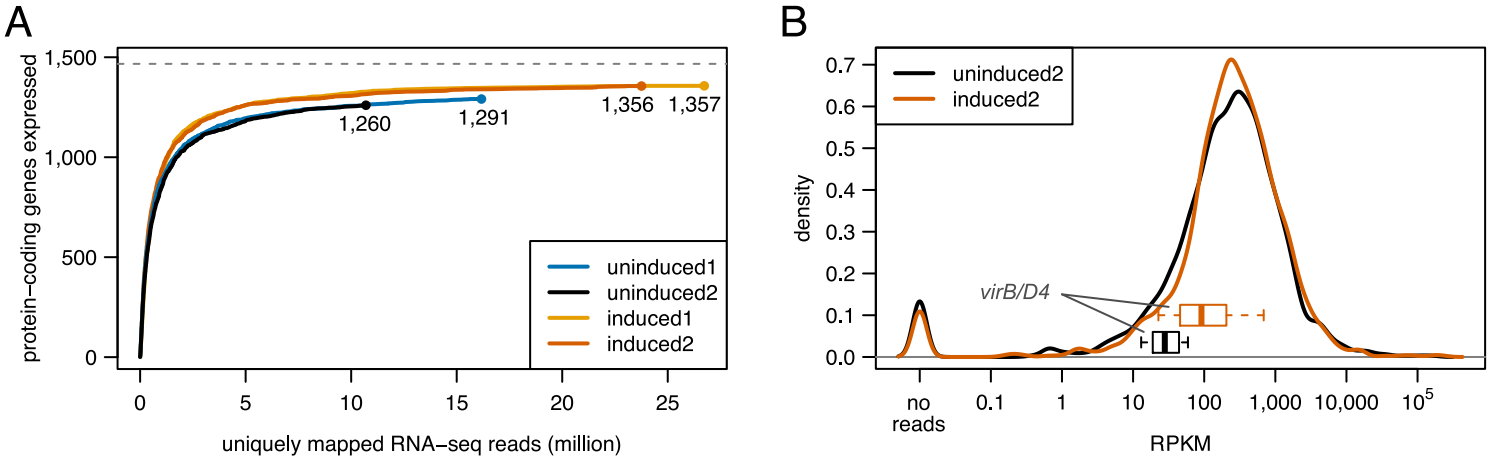
Venter, E., Smith, R.D., and Payne, S.H. 2011. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS ONE* 6: e27587.

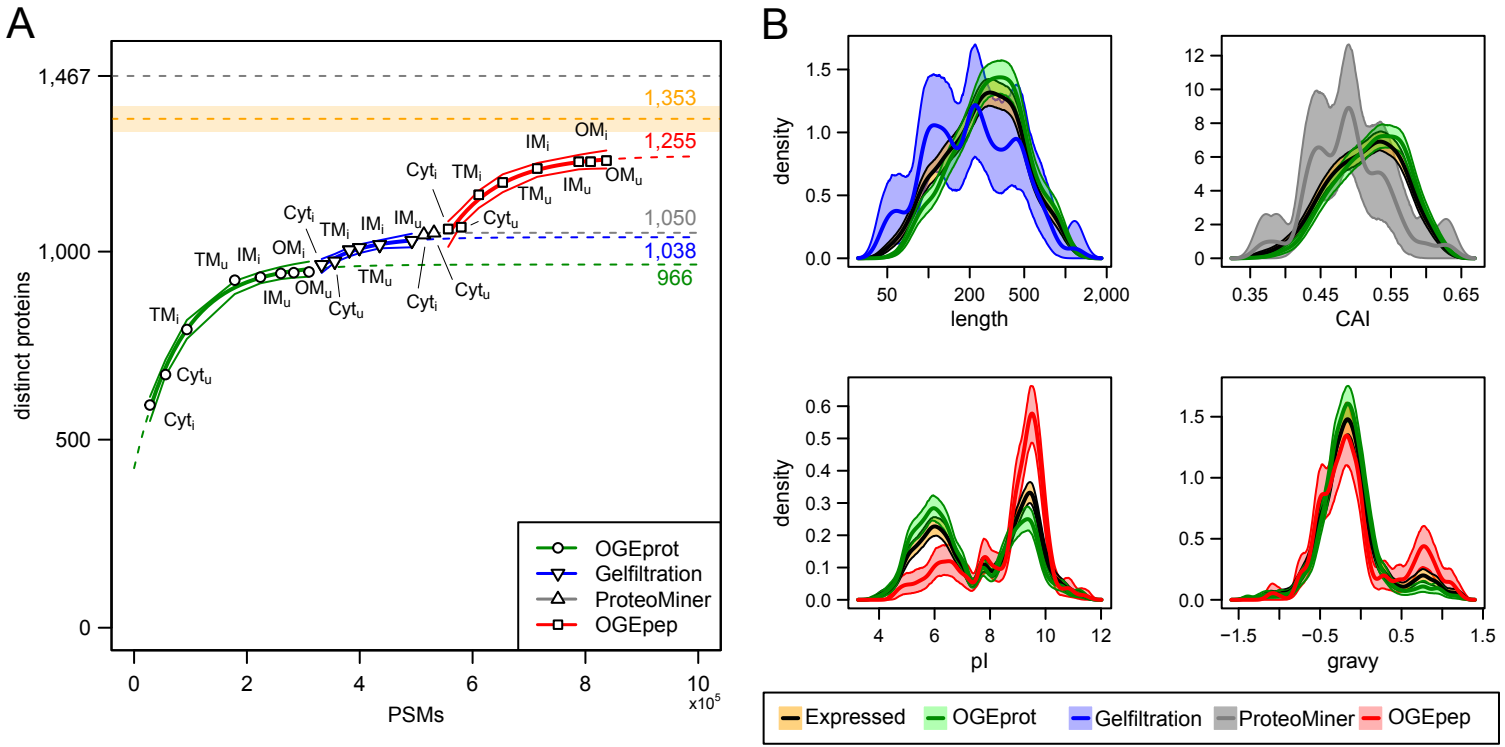
Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63.

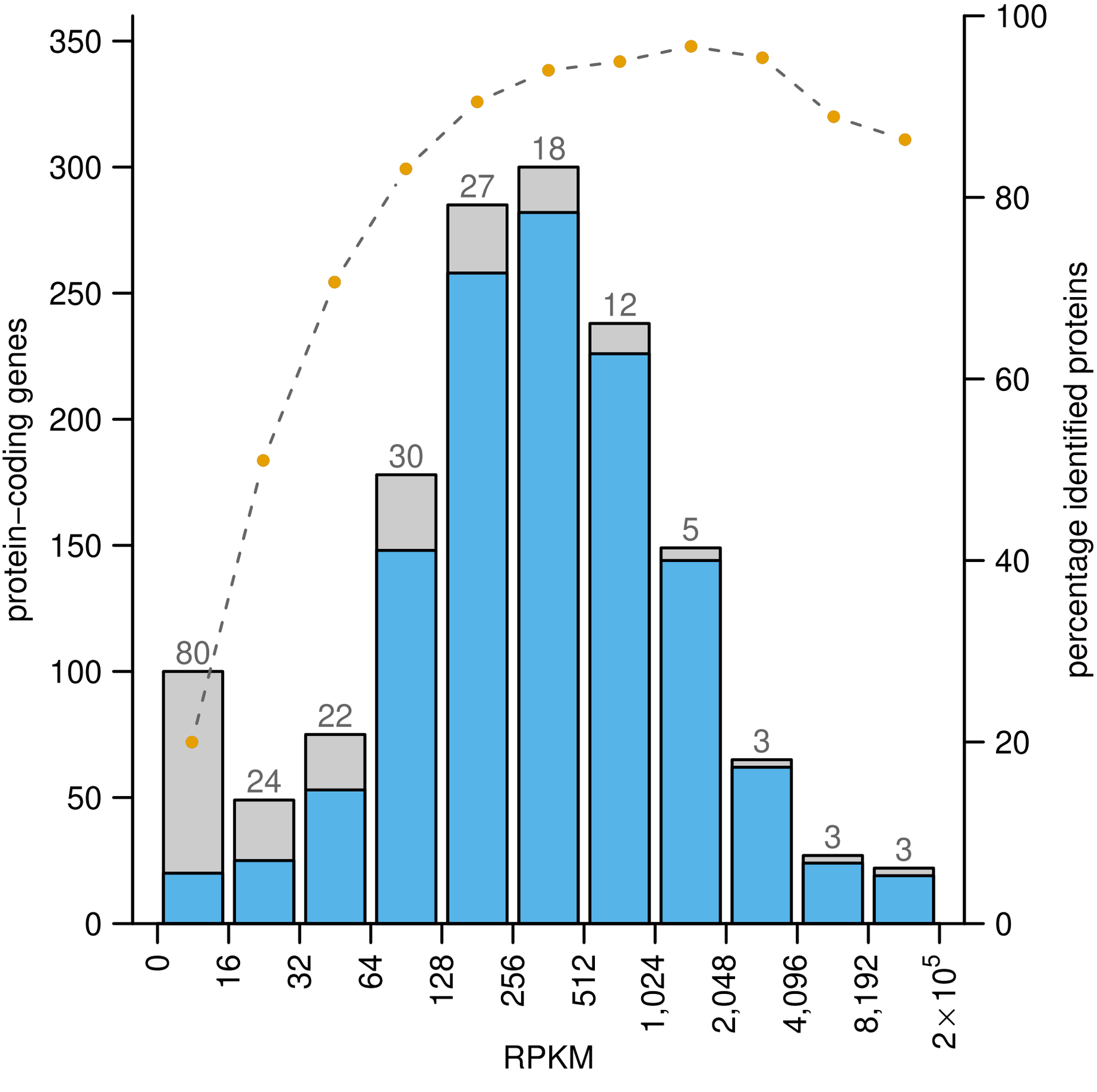
Westermann, A., J., Gorski, S., A., and Vogel, J. 2012. Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology* 10: 618-630.

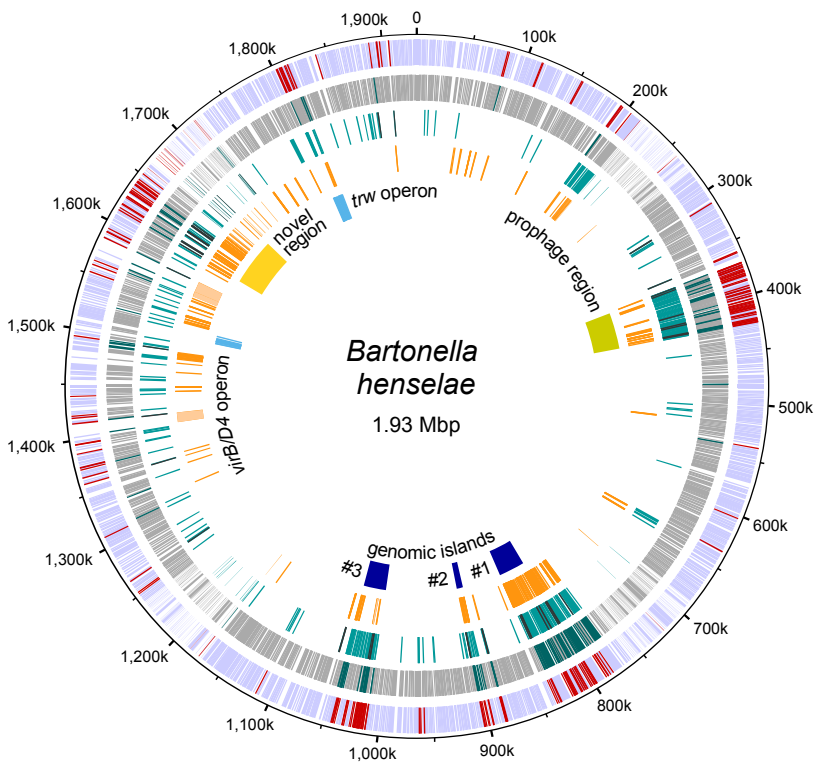
Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J. et al. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26: 1608-1615.


A
total RNA
saturated RNA-seq
induced
uninduced
proteins
subcellular fractionation:
Cyt
TM
IM
OM
B
under-represented proteome
statistical analysis
proteins
exp.
targeted experiments
identify proteins
C
complete expressed proteome
-
+
membrane proteome
subset correlations
differential expression
orthology analysis

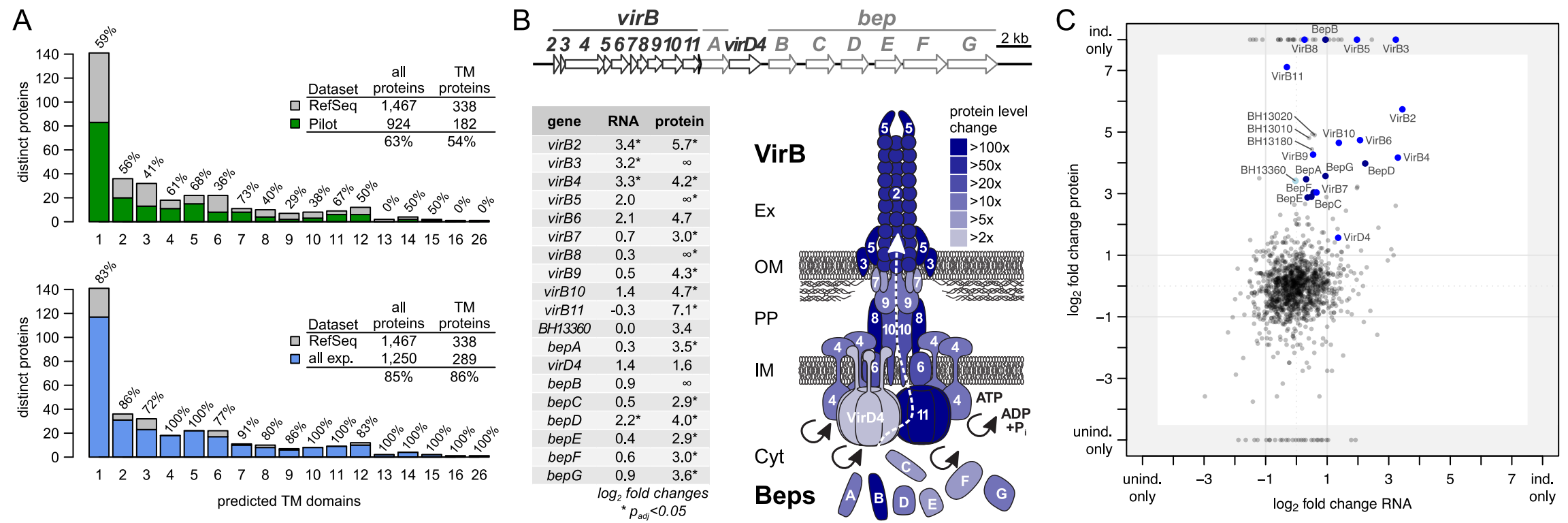

A
protein−coding genes expressed
1,500
1,000
500
0
0
5
10
15
20
25
uniquely mapped RNA−seq reads (million)
1,260
1,291
1,356
1,357
uninduced1
uninduced2
induced1
induced2
B
density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
uninduced2
induced2
virB/D4
no reads
0.1
1
10
100
1,000
10,000
10^5
RPKM

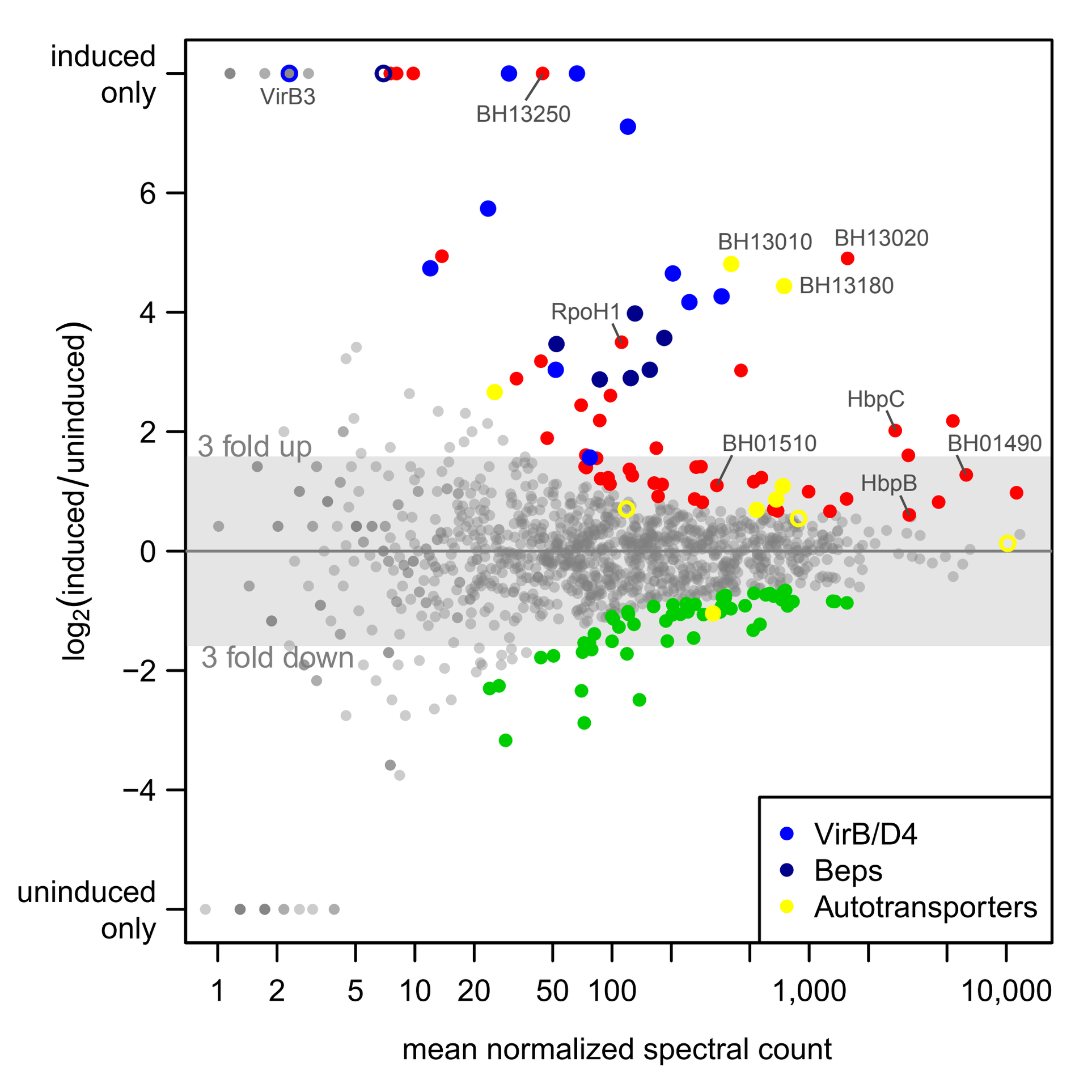

A
1,467
1,353
1,255
1,050
1,038
966
1,000
500
0
distinct proteins
0
2
4
6
8
10
x10^5
PSMs
OGEprot
Gelfiltration
ProteoMiner
OGEpep
B
density
length
CAI
pI
gravy
Expressed
OGEprot
Gelfiltration
ProteoMiner
OGEpep


protein−coding genes
percentage identified proteins
RPKM
0
50
100
150
200
250
300
350
0
20
40
60
80
100
80
24
22
30
27
18
12
5
3
3
3
0
16
32
64
128
256
512
1,024
2,048
4,096
8,192
2×10^5



	proteins	
	seen	not seen
1,488 protein-coding genes	1,261	227
198 genes in gen. islands or prophage region	89	109*
59 genes in novel genomic region	28	31*
395 genes without ortholog	208	187*
55 genes without functional annotation	3	52*


A
distinct proteins
Dataset
all proteins
TM proteins
RefSeq 1,467 338
Pilot 924 182
63% 54%
59% 56% 41% 61% 68% 36% 73% 40% 29% 38% 67% 50% 0% 50% 50% 0% 0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 26
RefSeq 1,467 338
all exp. 1,250 289
85% 86%
83% 86% 72% 100% 100% 77% 91% 80% 86% 100% 100% 83% 100% 100% 100% 100% 100%
predicted TM domains
B
virB
bep
2 3 4 5 6 7 8 9 10 11 A virD4 B C D E F G
2 kb
gene RNA protein
virB2 3.4* 5.7*
virB3 3.2* ∞
virB4 3.3* 4.2*
virB5 2.0 ∞*
virB6 2.1 4.7
virB7 0.7 3.0*
virB8 0.3 ∞*
virB9 0.5 4.3*
virB10 1.4 4.7*
virB11 -0.3 7.1*
BH13360 0.0 3.4
bepA 0.3 3.5*
virD4 1.4 1.6
bepB 0.9 ∞
bepC 0.5 2.9*
bepD 2.2* 4.0*
bepE 0.4 2.9*
bepF 0.6 3.0*
bepG 0.9 3.6*
log2 fold changes
* padj<0.05
VirB
Ex
OM
PP
IM
Cyt
Beps
protein level change
>100x
>50x
>20x
>10x
>5x
>2x
ATP
ADP +Pi
VirD4
A B C D E F G
C
ind. only
unind. only
log2 fold change protein
log2 fold change RNA
BepB
VirB8
VirB5
VirB3
VirB11
VirB2
BH13020
BH13010
BH13180
VirB10
VirB6
VirB9
VirB4
BH13360
BepA
BepG
BepD
BepF
VirB7
BepE
BepC
VirD4


induced only
uninduced only
log2(induced/uninduced)
6
4
2
0
−2
−4
3 fold up
3 fold down
VirB3
BH13250
BH13010
BH13020
BH13180
RpoH1
HbpC
BH01510
BH01490
HbpB
VirB/D4
Beps
Autotransporters
1
2
5
10
20
50
100
1,000
10,000
mean normalized spectral count